LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-235951

Biosecurity Techbase FY07 Final Report

S. N. Gardner, P. L. Williams

October 30, 2007

Disclaimer

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Biosecurity Techbase FY07
Final Report

Shea N. Gardner, Peter L. Williams
Computations Directorate, BACE Division, Lawrence Livermore National Laboratory
POC: Peg Folta

Abstract
This tech base award has close links with the Viral Identification Characterization Initiative (VICI) ER LDRD. The tech base extends developed code to enable a capability for biodefense, biosurveillance, and clinical diagnostics. The code enables the design of signatures to detect and discover viruses, without relying on prior assumptions as to the species of virus present. This approach for primer and signature design contrasts with more traditional PCR approaches, in which a major weakness is the unlikelihood of viral discovery or detection of unanticipated species. There were three crucial areas of the project that were not research and development, so could not be funded under the ER LDRD, but were a reduction to practice of the existing VICI algorithm that were necessary for the success of overall computational project goals. These areas, funded by the 2007 Tech Base award, were 1) improvement of the code developed under the VICI LDRD by incorporating T_m and free energy predictions using Unafold; 2) porting of code developed on the kpath Sun Solaris cluster to the Yana and Zeus LC machines; and 3) application of that code to perform large numbers of simulations to determine parameter effects.

Background
Bioinformatics has been essential in securing LLNL's national leadership in pathogen detection and diagnostics. We have developed the computational algorithms and a high throughput infrastructure to guide the development of assays and instrumentation to successfully detect a number of deadly pathogens. Currently, fielded technologies require a diagnostic test per known pathogen. Our next steps at advancing the field are aimed at identification of any of a number of traditional, enhanced, emerging, and advanced pathogens within a single diagnostic, by means of multiplexed assays. Extending the computational infrastructure will increase our capabilities to meet the nation's need. Optimizing our current computational algorithms will enable the development of multiplexed assays and validate our ability to identify know and unknown viruses. We will extend our bioinformatics capabilities to guide extensive experimental testing, specifically targeted at optimization of PCR conditions for DNA amplification using short primers.

Work Performed
The overall goal of VICI is to develop methods and a platform for viral detection and discovery. The bioinformatic component involves prediction of sets of highly conserved universal primers capable of specific amplification of both well-characterized and

unsequenced novel viruses. With LDRD funding, we developed algorithms and software to predict univeral viral primer sets. However, we wished to incorporate an accurate, fast method for T_m and free energy calculation to predict and avoid primer dimerization, homodimerization, and hairpin formation. This would enable us to predict primer sets to better function in multiplex by selecting primers unlikely to participate in undesired reactions. Thus, 2007 Tech Base funding was used to modify the VICI code to call on the LLNL version of Unafold for T_m and free energy predictions. This LLNL version of Unafold was developed under 2006 Tech Base award, which enabled us to speed up the existing (open source) software by an order of magnitude as a result of more efficient data structures and memory access.

Once we had the improved T_m and free energy predictor included in the VICI software, we needed to run it many times with various input data and parameter settings. For this, we ported the software to the LC clusters Yana and Zeus, adding restart capability due to the process time limits imposed on those systems. We also installed all the underlying perl module libraries, including the LLNL version of Unafold, on Yana and Zeus.

Third, we ran simulations using the VICI code to determine how primer length affects multiplex demands and the ability to detect emergent viruses. We asked how many primers would be required such that at least one fragment in a detectable length range from each of the > 35K virus sequences currently available would be amplified (counting each complete segment or genome). Throughout, we will use the term "detected" to mean that at least one amplicon in a pre-specified length range is generated. The algorithm applies to both single and double stranded RNA and DNA organisms because of a reverse-transcriptase step, which creates double stranded DNA based on the template, prior to the PCR reaction.

These calculations were performed to identify the empirical approaches for PCR amplification and detection of novel viruses that might be feasible, and to illustrate the challenges posed by finding universal viral primers. The number of primers in a universal set increases with the number of available genomes and their diversity, so we anticipate that as new genomes become available, we will need to update the universal primer sets by re-running this code on the ever-growing viral sequence database.

Results

With 10-mer primers, 1974 primers are required to detect all viral genomes, but this number drops by over half, down to 764 primers, if it is sufficient to detect only 90% of viral targets (Figure 1). For 9-mers, 1754 and 525 primers are required to detect 100% or 90%, respectively. Using longer primers such as 15-mers requires 3318 primers for 100% or 1246 primers for 90%. With primers as short as 7-mers, it is not possible to amplify all target genomes, since many do not contain 7-mers in the required T_m range within an PCR-amplifiable distance of one another. Thermodynamic constrains on the allowable primer parameters (e.g. free energies, Tm's) increases the number of primers required to amplify all viruses (data not shown), because some highly conserved primers may be eliminated, and in some cases results in an inability to find primers for all targets, as for 7-mers. For example, by removing the primer Tm and GC% constraints, the number of

primers in a universal set for all viruses can drop by up to 80% or more: for 9-mers, this number is 623 primers, compared to ~1700 primers if a relatively high T_m is required (64% fewer primers). This prediction has driven labwork to develop protocols for specific PCR amplification using short, low-T_m primers.

The increase in sequence data between 2004 and 2007 require approximately 700 more 10-mer primers to amplify all sequenced viruses in 2007 compared to 2004 (Figure 2). While the increase in the number of sequences used between the two dates was only ~15%, the number of primers required increased by 48%, illustrating the substantial increase in diversity represented by the additional sequence data. Using 10 to 15-mer primers generated based on the 2004 sequence data, simulations indicate that only ~35% of the genomes sequenced in 2007 would have been detected (Figure 3). Shorter primers increase this fraction.

The ability to detect/discover unknown viruses using viral universal primer sets depends strongly on primer size (Figure 4). With 7-9-mers, our simulations indicate that the majority of newly emerged viruses would be detected. With 10-mers, more than half would have been detected. With primers longer than 10 bases, however, our simulations indicate that it is more likely that these newly emerged viruses would not have been detected. The amount of sequence data used to generate the universal primer sets also plays a role, such that with 9-10-mers, about 20% more unknowns could be detected (in the absence of sequence information for those unknowns and their subsequently discovered near neighbors) using primer sets generated in 2007 than in 2004.

Discussion and Conclusions

Using bioinformatics we have analyzed some of the requirements and challenges that must be surmounted in order to develop universal PCR for viral detection and discovery. One major factor that will influence the success of this approach is the available viral sequence data used to create universal primer sets. This problem is compounded by the huge predicted viral diversity on earth, the rapid evolution of RNA and phage viruses, and the lack of cultureable viruses to characterize. As our simulations indicate, the modest increase in the total number of sequences available between 2004 and 2007 made a dramatic difference in the size of the universal primer sets, and 65% of the genomes available in 2007 would have gone undetected using primer sets generated in 2004. It is expected, therefore, that these signatures will evolve following the process defined above at regular intervals, as our sequence databases grow. It is the viruses in tail that are divergent relative to other sequenced viruses which each require a unique primer pair, that drive up the size of a universal primer set. This depends on sequence availability and bias, however, since many species are represented by only one sequence, and some families like Filoviridae by only two genera: Ebola-like and Marburg viruses. The shape of the tail in Figure 1 and the size of the universal set is likely to change as more viruses are sequenced.

Despite these challenges, we have computationally demonstrated the ability of a universal primer set to detect newly emerging viruses such as Ebola, SARS, Nipah, and Hendra viruses. The large size of the universal primer set, however, for multiplexing capabilities,

and blurs the distinction between specific amplfication and random amplification. To our knowledge, no one has studied how many unique primer sequences are actually contained in random primer kits available for purchase. Nor has anyone determined what fraction of those random primer sequences actually can participate in priming reactions, possibly failing as primers due to low T_m's, hairpin, or dimer formation.

Another major challenge to this approach is to develop a protocol for specific (as opposed to random) amplification using short primers. Since primers are short, it is essential to purify the sample, excluding as much as possible contaminating nucleic acids from eukaryotic and prokaryotic sources, since simulations indicate (not shown here) that sets of short multiplex primers will generate hundreds or thousands of amplicons from a human or typical bacterial genome that would be difficult to discriminate by sequencing or banding pattern. The assay module of the VDP LDRD has been successful in developing such a protocol for viruses using material from a standard DNA extraction without special measures to remove contaminating nucleic acids (Hiddessen et al, in prep.).

In conclusion, we have developed a set of universal primers for viral detection and discovery. Due to high viral diversity, short primer, multiplexed PCR will be necessary in order to detect novel viruses. These bioinformatic simulations have driven lab experimentation, since the size of the multiplex will depend on capabilities in the laboratory to amplify using short priming sequences with a wide range of T_m's.

Figure 1: Percent of viral genomes detected vs number of primers required, based on sequence data as of April 25, 2007, for primers of size 6-15 bases.

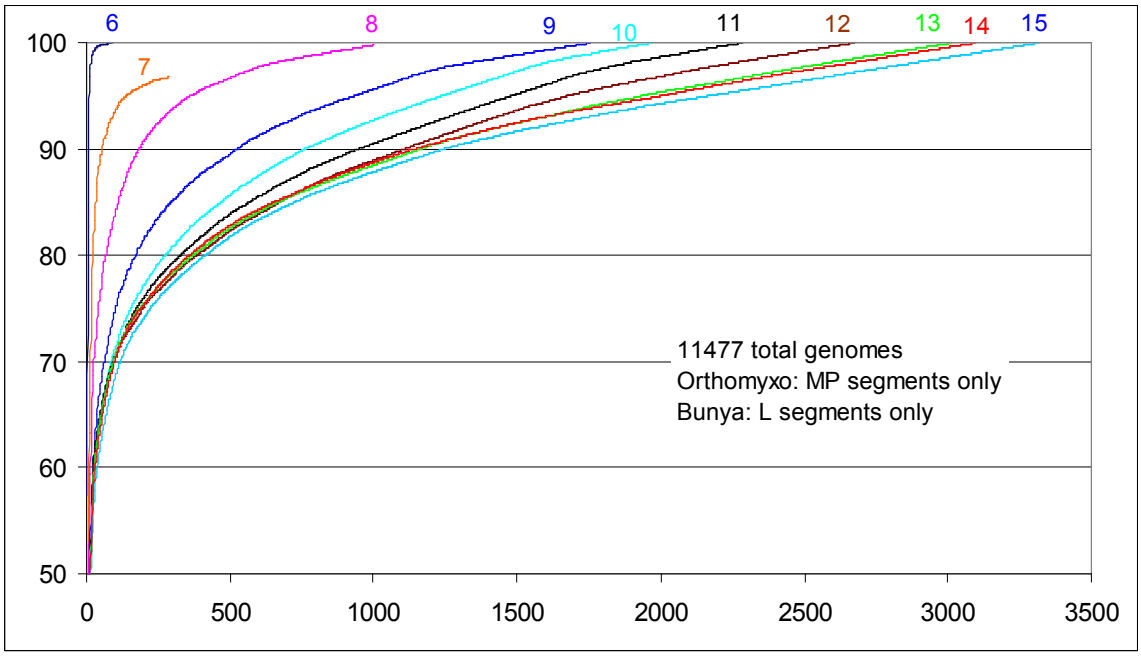

Figure 2: Percent of viral genomes detected vs number of primers required, based on sequence data as of January 1, 2004, for primers of size 7-15 bases.

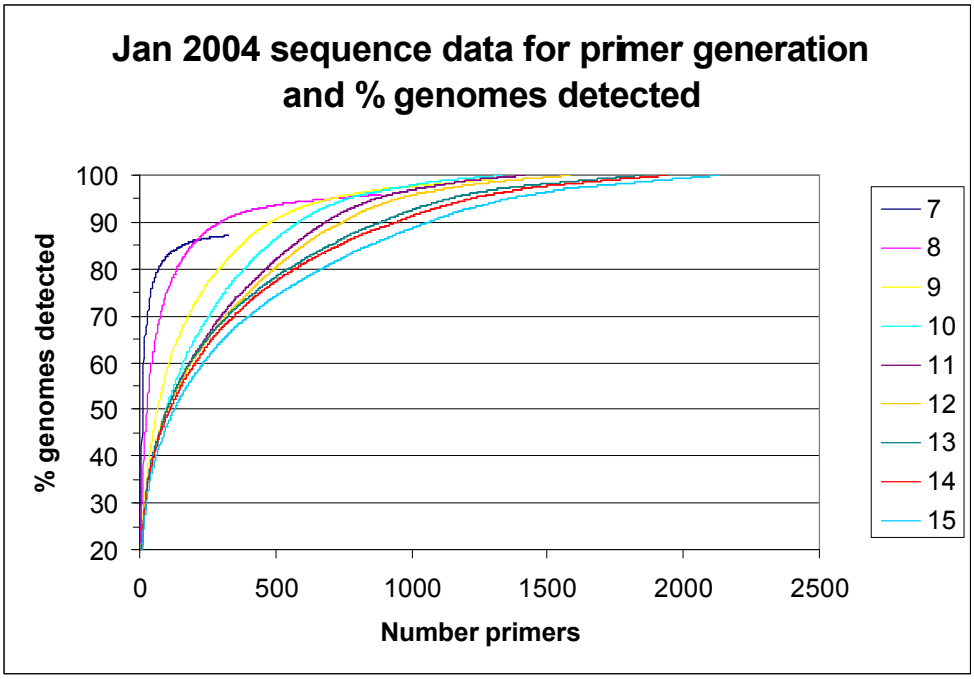

Figure 3: Fraction of viral genomes available on April 25, 2007 that would have been detected using primer sets developed based on the sequence data available on January 1, 2004.

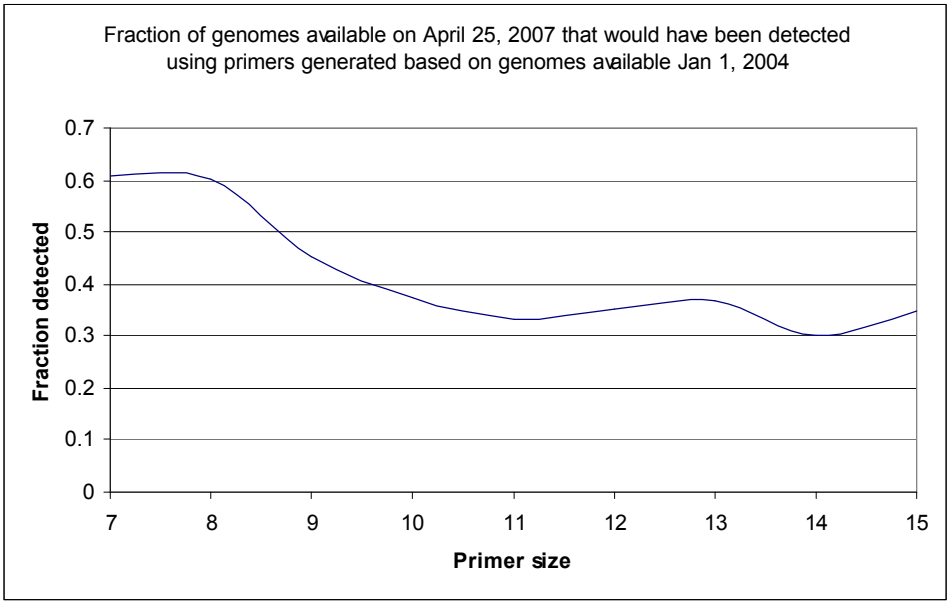

Figure 4: Percent of newly emerged ("unknown" viruses) that would have been detected based on sequence data available as of January 1, 2004 or April 25, 2007, as a function of primer length.